贝叶斯因子及其在 JASP 中的实现

胡传鹏 [1, 2*]; 孔祥祯 [3]; Eric-Jan Wagenmakers [4]; Alexander Ly [4]; 彭凯平 [1*]

1 清华大学心理学系, 北京 100084

2 Neuroimaging Center, Johannes Gutenberg University Medical Center, 55131 Mainz, Germany

3 Language and Genetics Department, Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands

4 Department of Psychological Methods, University of Amsterdam, 1018 VZ Amsterdam, The Netherlands,

chinaXiv:201709.00120v4

注：本文已经被《心理科学进展》接收，并将发表于《心理科学进展》2018 年 6 月刊上，见（http://journal.psych.ac.cn/xlkxjz/CN/Y2018/V26/I6/951）。

引用本文：

胡传鹏, 孔祥祯, Eric-Jan Wagenmakers, Alexander Ly, 彭凯平. 贝叶斯因子及其在 JASP 中的实现[J]. 心理科学进展, 2018, 26(6): 951-965. doi:10.3724/SP.J.1042.2018.00951

HU Chuan-Peng, KONG Xiang-Zhen, Eric-Jan WAGENMAKERS, Alexander LY, PENG Kaiping. The Bayes factor and its implementation in JASP: A practical primer. Advances in Psychological Science, 2018, 26(6): 951-965. doi:10.3724/SP.J.1042.2018.00951

收稿日期: 2017-10-10

* 通讯作者: 胡传鹏 E-mail: hcp4715@hotmail.com;
彭凯平 E-mail: pengkp@mail.tsinghua.edu.cn

chinaXiv:201709.00120v4

摘 要　　统计推断在科学研究中起到关键作用，然而当前科研中最常用的经典统计方法——零假设检验（Null hypothesis significance test, NHST）却因难以理解而被部分研究者误用或滥用。有研究者提出使用贝叶斯因子（Bayes factor）作为一种替代和（或）补充的统计方法。贝叶斯因子是贝叶斯统计中用来进行模型比较和假设检验的重要方法，其可以解读为对零假设 H_0 或者备择假设 H_1 的支持程度。其与 NHST 相比有如下优势：同时考虑 H_0 和 H_1 并可以用来支持 H_0、不“严重”地倾向于反对 H_0、可以监控证据强度的变化以及不受抽样计划的影响。目前，贝叶斯因子能够很便捷地通过开放的统计软件 JASP 实现，本文以贝叶斯 t 检验进行示范。贝叶斯因子的使用对心理学研究者来说具有重要的意义，但使用时需要注意先验分布选择的合理性以及保持数据分析过程的透明与公开。

关键词　　贝叶斯因子, 贝叶斯学派, 频率学派, 假设检验, JASP

自 20 世纪以来，统计推断在科学研究中起到越来越重要的作用(Salsburg, 2001)，科学研究结论的正确性也越来越依赖于统计推断的正确应用。目前，使用最为广泛的统计推断方法是零假设检验（Null hypothesis significance testing, NHST）(Wasserstein & Lazar, 2016)。然而，与 NHST 在各个领域中广泛使用相伴的是研究者对 NHST 及 *p* 值的误解和盲目使用(Gigerenzer, 2004; Greenland et al., 2016; Ziliak & McCloskey, 2008; 胡传鹏等, 2016; 骆大森, 2017)，因此带来一些消极的后果。例如，*p* 值被用来支持不合理且无法重复的研究结果(如, Bem, 2011)，引起了关于 NHST 是否适合于科学研究的争论(Miller, 2011)。在这个背景之下，有研究者推荐使用贝叶斯因子替代 NHST(Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; 钟建军, Dienes, Z., 陈中永, 2017)。

贝叶斯因子（Bayes factor）是贝叶斯统计（Bayesian statistics）中用来进行模型比较和假设检验的方法。在假设检验中，其代表的是当前数据对零假设与备择假设支持的强度之间的比率。正如下一节将要详述的，贝叶斯因子能够量化地反映当前数据对各个假设支持的程度，因此可能更加适用于科研中的假设检验。但由于贝叶斯因子的统计原理及实现相对复杂，其在各个学科的研究中并未获得广泛应用。

近年来，随着计算机运算能力的大大提升，贝叶斯统计在计算机等领域获得了巨大的成功，如(Zhu, Chen, Hu, & Zhang, 2017)。贝叶斯统计的工具迅速发展，如 WinBUGs (Lunn, Spiegelhalter, Thomas, & Best, 2009)、JAGS (Plummer, 2003)、Stan (Carpenter et al., 2017)和 python 语言的工具包 PyMC3（http://docs.pymc.io/index.html）等。这些软件和工具包的出现，促进贝叶斯方法在各个研究领域中的使用(Depaoli & van de Schoot, 2017; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017)。在这些工具中，也出现了用于计算贝叶斯因子的工具，如 R 语言中的 BayesFactor（http://bayesfactorpcl.r-forge.r-project.org/）。在心理学及

chinaXiv:201709.00120v4

相关领域，最近有不少研究者试图引入贝叶斯统计的方法(Dienes, 2008, 2011, 2014; Hoijtink, 2011; Klugkist, Laudy, & Hoijtink, 2005; Kruschke, 2014; Masson, 2011; Morey & Rouder, 2011; Mulder et al., 2009; Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Vanpaemel, 2010; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010)。在心理学(Open Science Collaboration, 2015; 胡传鹏等, 2016)、神经成像研究(Chen, Lu, & Yan, 2018; Zuo & Xing, 2014)等领域出现“重复危机”的背景之下，使用合理的统计方法显得更加迫切。但对于不少心理学及相关领域的研究者来说，使用 R 语言或其他计算机语言进行贝叶斯因子计算仍然较为困难。为解决这一障碍，研究者们开发了与商业统计软件 SPSS 具有相似图形界面的统计工具 JASP (https://jasp-stats.org/, JASP team 2017) (JASP Team, 2017; Marsman & Wagenmakers, 2016a; Wagenmakers, Love, et al., 2017; Wagenmakers, Marsman, et al., 2017)，简化了贝叶斯因子的计算。

本文旨在为向心理学及相关学科的研究者介绍贝叶斯因子及其使用。首先本文将介绍贝叶斯因子的原理，及其相对于传统假设检验中 p 值的优势；再以独立样本 t 检验为例，介绍了如何使用 JASP 计算贝叶斯因子，以及如何解读和报告其结果。在此基础上，讨论了贝叶斯因子的应用价值及其不足。

1 贝叶斯因子的原理

贝叶斯因子是贝叶斯统计在假设检验上的应用，因此要理解贝叶斯因子，首先需要理解贝叶斯统计的原理。

1.1 贝叶斯统计简介

贝叶斯学派（Bayesian statistics）与频率学派（Frequentist statistics）是统计学中主要的两个学派，其核心的差异在于他们对于概率（probability）所代表的意义有着不一样的解读。对于

chinaXiv:201709.00120v4

频率学派而言，概率是通过无数次重复抽样中频率（frequency）的预期值。与之相反，贝叶斯学派则认为，概率是对一件事情的相信程度，从 0 到 1 表示人们基于所获得的信息，在多大程度上相信某件事情是真的。由于不同人对同一事件的相信程度可能不同，因此，贝叶斯学派的概率是具有主观性。但贝叶斯学派的概率却不是任意的：人们通过合理的方式，不断获取并更新已知信息，可以最终消除主观性，从而达成一致。

正由于频率学派将概率看作长期行为表现的结果，要理解频率学派的概率，通常需要假想尚未发生的事件。例如，在 NHST 框架之下，p 值的意义是假定 H_0 为真的情况下，出现当前结果及比当前结果更加极端结果的概率。换句话说，p 值表达的意思是：假如 H_0 为真，如果采用完全相同的条件，无数次地重复当前实验，这些实验中将有多大比例会出现当前结果模式或者比当前结果模式更极端的模式。因此，p 值的意义暗含一个重要的假设：我们能够无数次地重复试验。但研究者却经常忽略这种无数次重复相同试验的假定，误认为 p 值是单次检验中拒绝零假设时犯错误的概率(Greenland, et al., 2016)。这种对 NHST 的误解，恰好是带有贝叶斯统计色彩，即根据当前的数据计算某个模型正确或者错误的概率。

与频率学派统计不同，贝叶斯统计最大的特点之一在于：它考虑了不同可能性对于个体来说的可信度（credibility）(Kruschke, 2014)。而通过不断获得的数据，人们可以改变对不同可能性的相应程度。这种思维方式与人们在日常生活中的经验非常相似：当我们不断地获得支持某个观点的证据时，我们会更加相信该观点。

虽然贝叶斯统计对概率的理解与频率学派不同，但是其对概率的计算却严格依照概率的基本原则：加法原则与乘法原则。贝叶斯统计中最核心的贝叶斯法则（Bayes rule），也是根据简单的加法原则与乘法原则推导而来。依据概率的乘法原则，随机事件 A 与随机事件 B 同时发生的概率为：

$$p(\mathrm{A} \cap \mathrm{B}) = p(A|B) \times p(B) = p(B|A) \times p(A) \qquad \text{公式 1}$$

chinaXiv:201709.00120v4

公式 1 即为联合概率的公式，即 A 与 B 同时发生的概率。其意义为：A 与 B 的联合概率（$p(A\cap B)$）为，在 B 发生的条件下 A 发生的概率（$p(A|B)$）与 B 发生的的概率（$p(B)$）的乘积，也等于在 A 发生的条件下 B 发生的概率（$p(B|A)$）与 A 发生的概率（$p(A)$）的乘积。其中，$p(A|B)$和 $p(B|A)$均为条件概率（conditional probability），二者意义不同。

对公式 1 进行变换，即可以得到如下公式：

$$p(A|B) = \frac{p(A\cap B)}{p\ (B)} = \frac{p(B|A) \times p(A)}{p\ (B)} \qquad \text{公式 2}$$

公式 2 即为贝叶斯定理公式。其代表的意义是，如果我们要计算 B 发生的条件下 A 发生的概率（$p(A|B)$），可以通过使用 A 与 B 同时发生的概率（$p(A\cap B)$）除以 B 发生的概率（$p(B)$），也就等于在 A 发生的条件下 B 发生的概率，与 A 发生概率的乘积，再除以 B 发生的概率。公式 2 将两个条件概率联系起来，从而使得计算不同的条件概率成为可能。

在贝叶斯统计的框架之下，公式 2 可以看作是一次信息的更新。假定我们需要根据一次实验收集到的数据（data）来检验某个理论模型为真的可能性。以心理学研究中常用的零假设 H_0 为例，则可以将公式 2 改写如下：

$$p(H_0|data) = \frac{p(data|H_0) \times p(H_0)}{p\ (data)} \qquad \text{公式 3}$$

$p(H_0|\text{data})$表示数据更新之后理论模型 H_0 正确的概率，即*后验概率*（posterior）；$p\ (H_0)$表示更新数据之前认为理论模型 H_0 正确的概率，即*先验概率*（prior）；而 $p(\text{data}|\ H_0)$则是在模型 H_0 之下，出现当前数据的概率，即*边缘似然性*（marginal likelihood）。由此可以看出，在贝叶斯统计之中，一次数据收集（实验）的主要功能在于帮助我们更新理论模型的可信度。

根据公式 3，我们可以使用数据对任意的模型为真的概率进行更新。在假设检验中，我们可以根据观测数据同时对零假设（理论模型 H_0）和备择假设（理论模型 H_1）的可信度进行更新(分别见公式 3 和公式 4)，得到它们更新的后验概率。

$$p(H_1|data) = \frac{p(data|H_1) \times p(H_1)}{p\ (data)} \qquad \text{公式 4}$$

chinaXiv:201709.00120v4

得到 H_0 和 H_1 的后验概率后，可能对两者进行比较，即公式 5：

$$\frac{p(H_1\,|data)}{p(H_0\,|data)} = \frac{p(data|H_1)}{p(data|H_0)} \times \frac{p(H_1)}{p(H_0)} \qquad \text{公式 5}$$

其中，贝叶斯因子为：

$$\mathrm{BF}_{10} = \frac{p(data\,|\,H_1)}{p(data\,|\,H_0)} \qquad \text{公式 6}$$

在公式 6 中，BF_{10} 下标的 1 代表的是 H_1，0 代表的是 H_0，因此，BF_{10} 即代表的是 H_1 与 H_0 对比的贝叶斯因子，而 BF_{01} 则代表的是 H_0 与 H_1 对比的贝叶斯因子。例如，$\mathrm{BF}_{10} = 19$ 表示的是，在备择假设 H_1 为真条件下出现当前数据的可能性是虚无假设 H_0 条件下出现当前数据的可能性的 19 倍。从这个定义公式中可以看出，贝叶斯因子是体现了当前数据将先验概率更新为后验概率过程中的变化。

正是如此，贝叶斯因子与 NHST 回答了不同的问题。NHST 试图回答“假定我们已知两个变量的关系（如，两种条件没有差异），出现当前观测数据的模式或者更加极端模式的概率（p(more extreme > observed data | H_0)）有多大”的问题；而贝叶斯因子试图回答的是：“在当前数据更可能在哪个理论模型下出现”的问题。在假设检验中，贝叶斯因子具有一些 NHST 不具备的优势（见表 1），下一小节将对这些优势进行详细说明。

表 1. 假设检验中贝叶斯推断与传统 NHST 推断的比较

假设检验中的问题	贝叶斯因子	传统推理	参考文献
1. 同时考虑 H_0 和 H_1 的支持证据	√	×	10, 11
2. 可以用来支持 H_0	√	×	12, 13
3. 不“严重”地倾向于反对 H_0	√	×	14, 15, 16
4. 可以随着数据累积来监控证据的强度	√	×	17, 18
5. 不依赖于未知的或者不存在的抽样计划	√	×	19, 20

chinaXiv:201709.00120v4

注：10 = Jeffreys (1935)；11 = Jeffreys (1961); 12 = Rouder, et al. (2009); 13 = Wagenmakers (2007); 14 = Edwards (1965); 15 = Berger and Delampady (1987); 16 = Sellke, Bayarri, and Berger (2001); 17 = Edwards, Lindman, and Savage (1963); 18 = Rouder (2014); 19 = Berger and Berry (1988); 20 = Lindley (1993).

在 Jeffreys (1961)的基础上，Wagenmakers, et al. (2017)对贝叶斯因子的大小所代表的意义进行原则上的划分（见表 2）。但是这个划分仅是大致参考，不能严格对应，研究者需要根据具体的研究来判断贝叶斯因子的意义。

表 2. 贝叶斯因子决策标准

贝叶斯因子, BF_{10}	解释
> 100	极强的证据支持 H_1
30 - 100	非常强的证据支持 H_1
10 - 30	较强的证据支持 H_1
3 – 10	中等程度的证据支持 H_1
1 – 3	较弱的证据支持 H_1
1	没有证据
1/3 - 1	较弱的证据支持 H_0
1/10 – 1/3	中等程度的证据支持 H_0
1/30 – 1/10	较强的证据支持 H_0
1/100 – 1/30	非常强的证据支持 H_0
< 1/100	极强的证据支持 H_0

chinaXiv:201709.00120v4

1.2 备择假设的默认先验

由于贝叶斯因子中先验概率具有至关重要的作用，如何选择备择假设的先验分布变得尤其重要。其中一个较为合理的做法是，根据某问题的先前研究结果（如元分析得到的效应量）来设定备择假设的先验分布。但这种做法在很多情况下并不现实：首先根据范式的不同，效应量的可能分布不同；更重要地，由于许多研究本身具有一定的探索性，并没有先前研究结果作为指导。因此，更加常用的做法是使用一个综合的、标准化的先验。

例如，在贝叶斯 *t* 检验中，使用柯西分布（Cauchy distribution）作为备择假设的先验可能是比较合理的选择(Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, 2016a, 2016b; Rouder, et al., 2009)。与标准正态分布相比，柯西分布在 0 附近概率密度相对更小一些，因此其比标准的正态允许更多较大的效应（见图 1）；而与均匀分布（即效应量在所有值上的分布完全相同）相比，柯西分布更偏好零假设一些(Jeffreys, 1961; Rouder, et al., 2009)。因此，对于备择假设的先验分布，可以如下表示：

$$\delta \sim Cauchy(x_0 = 0, \gamma = 1)$$

chinaXiv:201709.00120v4

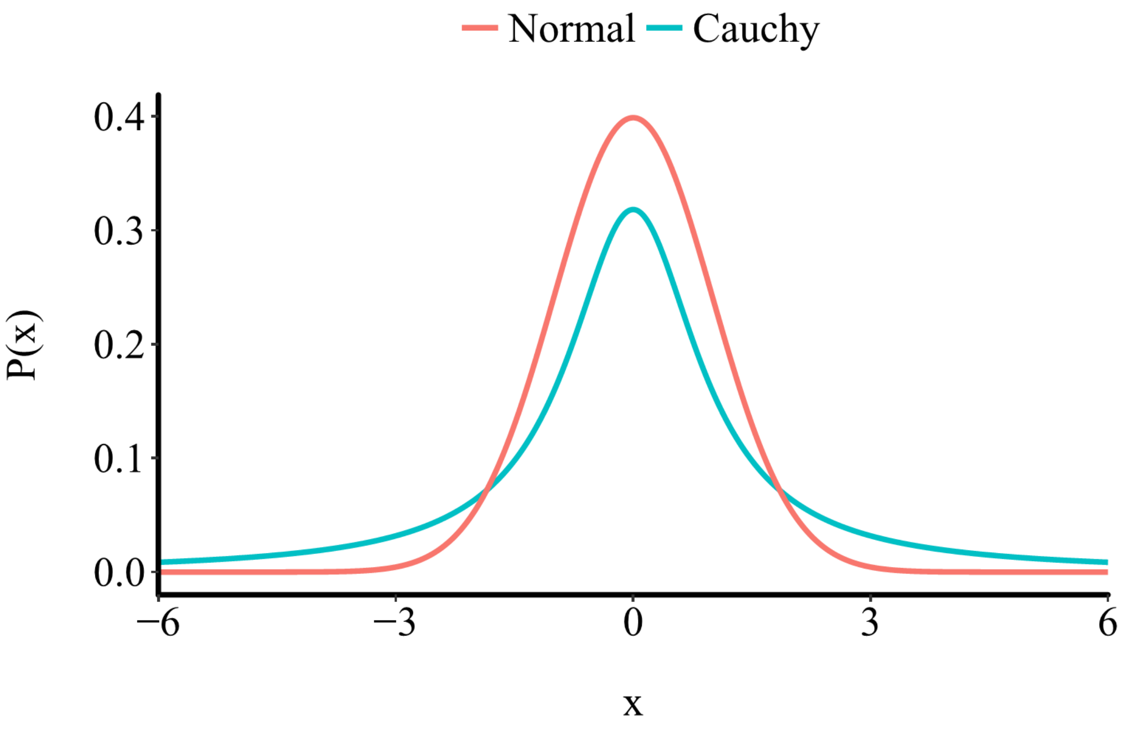

图 1 柯西分布与正态分布的对比

Jeffreys (1961)最早提出在贝叶斯因子中使用柯西分布作为先验来比较两样本的问题。最近研究者的进一步验证表明，柯西分布可以作为先验用于计算心理学研究中常用的贝叶斯因子分析，如 *t* 检验(Rouder, et al., 2009)、ANOVA(Rouder, et al., 2012)和相关分析(Ly, Marsman, & Wagenmakers, 2017; Ly, et al., 2016b)等。这些验证性的工作，为贝叶斯因子在心理学及相关学科研究中的应用打下了基础。

2 贝叶斯因子的优势

如前所述，在假设检验中，贝叶斯因子除了更加符合人们的直觉之外，还具有一些 NHST 所不具备的优势。这些优势可以总结为五个方面（见表 1）。以下将从这五个方面展开。

2.1 同时考虑 H_0 和 H_1

贝叶斯因子的计算同时考虑 H_0 和 H_1，并根据全部现有数据对 H_0 和 H_1 为真的先验概率进行更新，在此基础之上，比较在当前数据下哪个理论模型（H_0 和 H_1）更合理。这种思路与

chinaXiv:201709.00120v4

NHST 不同：在 NHST 框架之下，计算 p 值仅需要假定 H_0 为真，而对 H_1 不做任何假设，因此 p 值与 H_1 无关。NHST 的逻辑是，如果 H_0 为真的条件下，观察到当前数据出现的概率非常小，则拒绝 H_0，接受 H_1。这种情况下，NHST 忽略了一种可能性：当前数据下，H_1 为真的概率与 H_0 为真的概率相当或者更小(Wagenmakers, Verhagen, et al., 2017)。例如，在 Bem (2011) 中，H_0 是被试的反应不受到未来出现刺激的影响，H_1 是未来出现的刺激会影响到被试当前反应，即被试能够“预知”尚未出现的刺激。虽然采用 NHST 的逻辑 Bem (2011)得到了 $p < 0.05$ 的结果，即 H_0 为真时，得到当前数据的概率（$p(data|\mathrm{H0})$）很低，因此作者选择拒绝 H_0 而接受 H_1，认为被试能够预知未来出现的刺激。然而，研究者更关心的是，根据当前数据，我们能够得到某个模型/假设（如 H_1）为真的概率（$p(H_1|data)$），而非零假设 H_0 为真时得到当前数据的概率（$p(data|\ H_0)$）。在 Bem (2011)这个研究中，先验知识告诉我们 H_1 本身为真的概率可能非常低，在当前数据模式下，H_1 为真的可能性 $p(H_1|data)$极可能比 H_0 为真的可能性 $p(H_0|data)$更低(Rouder & Morey, 2011; Wagenmakers, et al., 2011)，但 NHST 却完全忽视了这一点。

2.2 可以用来支持 H_0

同样，由于贝叶斯因子同时量化当前数据对 H_0 和 H_1 各自的支持强度，其可以用来支持 H_0 (Dienes, 2014)。但是，在传统的 NHST 框架之下，假设检验仅在 H_0 为真的假设下进行，仅凭借小于显著性水平（比如 0.05 或 0.005）无法为 H_0 是否为真提供证据。比如，仅依据假设检验的结果 $p = 0.2$ 并不能推断*有证据表明没有效应*（evidence of absence）（除非结合样本量、效应量和统计效力 Power 做出综合判断）。

实际的研究中，能够对 H_0 提供量化的证据具有非常重要的意义(Gallistel, 2009; Rouder, et al., 2009)，它可以直观地让研究者区分出*有证据表明没有效应*（evidence of absence）和*没有证据表明有效应*（absence of evidence）这两种情况(Dienes, 2014)。具体来说，贝叶斯因子的

chinaXiv:201709.00120v4

结果有三种状态：（1）提供了支持 H_1 的证据（即*有证据表明有效应*)；（2）支持 H_0 的证据（即*有证据表明没有效应*)；或（3）证据对两者都不支持（*没有足够的证据表明有效应还是无效应*)。例如，贝叶斯因子 $BF_{01} = 15$ 表明观察到的数据出现在 H_0 为真情况下的可能性是在 H_1 为真情况下的可能性的 15 倍，表明***当前数据更加支持没有效应的假设 H*₀**。但是，假如 $BF_{01} = 1.5$，则说明观察到的数据出现在 H_0 为真情况下的可能性是在 H_1 为真情况下的可能性的 1.5 倍，则说明当前数据对于两个假设的支持程度相当，***没有足够的证据支持*** H_0 或者 H_1（见表 2 关于贝叶斯因子大小意义的建议）。

值得注意的是，不管是支持 H_1，还是支持 H_0，贝叶斯因子提供的证据是相对的，即相对于某个假设更支持另一个假设，因此可能存在第三个模型 H_2 比 H_1 和 H_0 均更接近真实情况，具有更高的后验概率。值得指出的是，最近有研究者在 NHST 框架之下发展出可以接受零假设的方法：等同性检验（Equivalence Test）。这种方法通过设定多个 H_0 来检验效应量是否与 0 没有差异，从而检验是否能接受 H_0(Lakens, 2017)。但等同性检验仍然使用了 p 值，无法提供对证据的直接测量(Schervish, 1996)。

2.3 不“严重”地倾向于反对 H_0

贝叶斯因子同时分别量化了当前数据对 H_0 和 H_1 支持的强度，其与传统 NHST 相比，其对 H_0 和 H_1 的支持更加均衡，从而其拒绝 H_0 的倾向也相对没有那么强烈。

在传统 NHST 假设之下，只要研究者能够收集足够多的数据，总能够得到 $p < 0.05$ 从而拒绝 H_0，与之相反的是，贝叶斯因子会随着数据的增加而逐渐趋于稳定（见后文 3.2 小节关于贝叶斯因子收敛的讨论）。对于同样的数据，p 值也似乎比贝叶斯因子对 H_0 的反对程度更强。例如，有研究者分析了美国总统选举中候选人的身高与当选之间的关系，对相关系数进行显著性检验之后发现 $r = 0.39$，$p = 0.007$ (Stulp, Buunk, Verhulst, & Pollet, 2013)。如果使用贝叶斯

chinaXiv:201709.00120v4

因子分析，则会得到 BF_{10} ＝ 6.33(Wagenmakers, Marsman, et al., 2017)。虽然两种方法大致上支持了同样的结论（即拒绝 H_0 与中等程度的证据支持 H_1），但是从 p 值上看，似乎表明拒绝 H_0 的证据很强，而贝叶斯因子得到的支持则是有保留的。Wetzels et al. (2011)比较了 855 个 t 检验的结果，发现虽然大部分的情况下 p 值与贝叶斯因子在结论上的方向一致，但是贝叶斯因子相对来说更加谨慎：p 值在 0.01 与 0.05 之间的统计显著结果，其对应的贝叶斯因子只表明有非常弱的证据。对传统 p 值的贝叶斯解读，详见(Johnson, 2013; Marsman & Wagenmakers, 2016b)。

2.4 可以监控证据的强度变化

计算贝叶斯因子时，可以根据数据来更新对 H_0 和 H_1 支持的程度，因此，随着新数据的出现，可以不断对不同假设的支持程度进行更新。在贝叶斯框架之下，贝叶斯因子的计算与解读均不需要假定存在无数的重复实验，而是按照似然性法则对贝叶斯因子进行更新，此外数据的出现顺序不会影响贝叶斯因子的解读(Rouder, 2014)。

贝叶斯统计的框架之下，不需要假定无数次重复试验，对贝叶斯因子的解读不会受到何时停止收集数据的影响(Rouder, 2014)。实际上，如果研究者们能够采用序列贝叶斯因子设计，在实验开始前提前设置贝叶斯因子的合理阈值（通常是 10，即较强的证据），则能够在实验中根据数据增加对后验概率进行更新，可以**在适当的时候**停止收集数据(Schlaifer & Raiffa, 1961; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017)。这种不受到停止规则影响的原则，对实际研究具有重要的意义，使得研究者能合理有效地收集数据。

2.5 不受抽样计划的影响

抽样计划指研究者根据数据分析的假设，在研究开始之前对样本选择以及数据收集过程进行计划以保证数据符合统计假设。例如，心理学实验中通常采用的随机抽样以及随机分配

chinaXiv:201709.00120v4

的做法。由于 NHST 的使用包含了一些潜在的假设，抽样计划（尤其是功效分析，power analysis）对于解读 p 值具有重要意义(Halsey, Curran-Everett, Vowler, & Drummond, 2015)。

但对于贝叶斯因子的解读，则不受到抽样计划的影响，原因在于贝叶斯因子的计算使用似然性原则(Berger & Wolpert, 1988)，其对数据的分析没有预先的假设。换句话说，即使研究者对数据收集的过程不清楚，仍能够计算和解读贝叶斯因子。这个特点对于分析自然情境中获得的数据非常实用。

仍然以上述的美国总统选举中候选人的身高与当选之间关系的研究为例，研究者发现 r = 0.39，p = 0.007 (Stulp, et al., 2013)。在 NHST 框架之下，要对 p 值进行合理的解读，我们必须假定实验者在总统选举之前已经计划好进行 46 次选举，并且在第 46 次选举后停止收集数据，并在此基础之上计算相关系数。如果不满足这些假设条件，p = 0.007 代表的意义很难解读。但很明显的是，这些假设是不成立的。

同样，这个例子还包含与停止规则（stopping rule，即什么条件下停止收集数据）相关的问题：在真实的生活中，美国的总统选举还会继续，数据会继续增加。如何分析未来的这些数据呢？如果每新增加一个数据均进行一次 NHST 分析，则会引起多重比较的问题，使得假阳性增加†。

与 NHST 不同，贝叶斯因子能够随着新数据不断地出现而不断地更新，从而能够分析实验室之外的真实数据，也能够对数据进行有意义的解读。从这个角度来讲，贝叶斯因子实时监控证据的优势与不受抽样计划影响的优势是相互关联的：这两个优势均是因为贝叶斯因子不依赖于研究者收集数据的意图。但是，正如我们在后面要提到的，虽然随着数据更新而更新贝叶斯因子不会影响到对其解读，但这种忽略假阳性的做法并不能避免假阳性的升高，研

† 对于频率主义的分析来说，多重比较是非独立的，校正的方法减少但不能消除一类错误。

chinaXiv:201709.00120v4

究者仍需要通过提前设置合理的阈值和（或）选择合适的先验来控制假阳性。

总之，贝叶斯因子以观察到的数据为条件，定量地分析当前数据对 H_0 和 H_1 提供的支持程度。通过实时地监控证据强度的变化，贝叶斯因子让研究者可以在收集数据的同时监控证据强度的变化。如果预先确定贝叶斯因子的停止阈值（比如 BF_{10} 大于 10 或者 BF_{10} 小于 1/10 时停止收集数据），研究者能够在证据足够充足停止收集数据。此外，即使缺乏数据收集计划信息的情况下，贝叶斯因子仍然能够从观测数据中得到证据来更加支持哪个假设。

3 使用 JASP 计算贝叶斯因子

由于贝叶斯因子的独特优势，因此很早就有研究者试图将其引入心理学的研究之中 (Edwards, et al., 1963)。但贝叶斯因子的计算在实际情况中随着数据类型和分析类型不同而变得更加复杂 (相关公式可以参考，Morey & Rouder, 2011; Rouder, et al., 2012; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017; Rouder, et al., 2009)。正是由于这个原因，贝叶斯因子在心理学的研究中一直受到很大的限制。最近，研究者利用 R 语言丰富的软件包，开发了可视化的统计工具 JASP (https://jasp-stats.org/)，该软件采用与 SPSS 类似的图形界面，让贝叶斯因子的计算变得更加容易实现，本小节将介绍 JASP 软件及其使用‡。

3.1 JASP 软件简介

JASP 是一个免费、开源的统计软件，其使用 R 语言的工具包进行数据处理，但其使用不需要安装 R。JASP 的长期目标是让所有人能够通过免费的统计软件进行最先进统计技术，尤其是贝叶斯因子。

JASP 是在心理学研究面临可重复危机的背景下开发的，其开发理念如下：第一，开源与免费，因为开源应该是科学研究的本质元素；第二，包容性，既包括贝叶斯分析，也包括

‡ 本小节内容部分来自于 Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2017). Bayesian Inference for Psychology. Part II: Example Applications with JASP. *Psychonomic Bulletin & Review*.

chinaXiv:201709.00120v4

NHST 分析方法，而且 NHST 分析方法中，增加了对效应量及其置信区间的输出(Cumming, 2014)；第三，简洁性，即 JASP 的基本软件中仅包括最常用的分析，而更高级的统计方法又可以通过插件模块进行补充；第四，友好的图形界面，例如，输出部分随着用户选择变量输入而实时更新，表格使用 APA 格式。同时，JASP 的使用递进式输出，即默认的结果输出是最简洁的，更多的结果输出可以由研究者自己进行定义。此外，为方便公开和分享分析过程，JASP 将输入的数据与输出结果保存于同一个后缀为.jasp 的文件之中，每个分析的结果均与相应的分析和变量数据相关联。这种结果与数据整合的文件可以与开放科学平台 Open science framework (OSF, https://osf.io/)兼容，从而做到数据与结果公开。

chinaXiv:201709.00120v4

3.2 贝叶斯因子分析在 JASP 的实现及其结果解读

目前，JASP 中可以实现多种实验设计的贝叶斯因子分析，包括单样本 *t* 检验、独立样本 *t* 检验、配对样本 *t* 检验、方差分析、重复测量的方差分析、ANCOVA 和相关分析。对于每一种分析，均提供了频率学派的方法和贝叶斯的方法。JASP 的贝叶斯因子分析中采用默认先验分布，但也可以修改。接下来本文将 Wagenmakers et al. (2015, https://osf.io/uszvx/)对 Topolinski 和 Sparenberg (2012)的重复实验数据为例进行分析，说明如何使用 JASP 进行独立样本 *t* 检验。其他常用贝叶斯因子分析，可以进一步参考 Wagenmakers 等人(2017)。

在 Topolinski 和 Sparenberg (2012)的第二个实验中，一组被试以顺时针方向拔动一个厨房用的钟，而另一组则以逆时针方向拨动。随后，被试填写一个评估经验开放性的问卷。他们的数据表明，被试顺时针转时比逆时针转的被试报告更高的对经验的开放性(Topolinski & Sparenberg, 2012)（但是见 Francis, 2013）。Wagenmakers 等人(2015)采用提前注册（preregistration）的方式对该研究进行重复，在实验开始前确定停止收集数据的标准：当支持某一个假设的贝叶斯因子达到 10 时即停止收集数据，或者每条件下达到 50 个样本后停止

收集数据。此外，预注册时采用单侧 t 检验的默认先验，即 $\gamma = 1$ 的柯西分布。而单侧的 t 检验的先验是只有正效应的柯西分布，即备择假设为 H_+ : Cauchy (0, 1)。

有研究者认为，默认先验分布 Cauchy (0, 1)是不现实的，因为在这个分布中，大的效应量占的比例太大（大于 1 的效应量在分布中占了 50%以上）；相反，另一些人觉得这个分布不现实是因为这个分布中，靠近 0 的效应量的比重太大，即效应量为 0 是最可能的值。一个避免这些问题的做法是减小柯西分布的参数 r。在 BayesFactor 工具包中，默认采用的

$$\gamma = \frac{1}{2}\sqrt{2} \approx 0.707$$

JASP 中对于单侧的 t 检验同样采用这个先验。r 减小意味着 H_1 和 H_0 相似，他们对观测数据的预测相似，更难得到支持 H_0 的强证据。

使用 JASP 可以对这批数据进行贝叶斯的独立样本 t 检验。首先用 JASP 打开数据（File → Examples → “Kitchen Rolls”，或者从 https://osf.io/9r423/下载后，点击 File → Open），然后在 T-tests 的面板中选择 “Bayesian Independent Samples T-test”。将显示如图 1 中间图所示的对话框。我们已经将“mean NEO”作为因变量（dependent variable），“Rotation”作为分组变量（grouping variable）。如图 2 中间所示，将 Cauchy 先验的宽度设置为 JASP 的默认值 $\gamma = 0.707$，同时勾选了“Prior and posterior”及其子选项的“Additional info”这两个选项，则得到如图 2 右侧所示的结果：与顺时针相比，逆时针对经验的开放性稍微高一些，这个结果的方向与 Topolinski 和 Sparenberg (2012)所假设的正好相反。图 2 右图下半部分中，实线为后验分布，虚线为先验分布。可以看到，大部分的后验概率是负值，其的中值是-0.13，95%的可信区间从-0.5 到 0.23。BF_{01} = 3.71，表明观察到的数据在 H_0 假设之下的可能性是在 H_1 假设之下可能性的 3.71 倍（我们选择了 BF_{01}，因为 BF_{01}＝3.71 相对于等价的 BF_{10}＝0.27 来说更好解释）。

chinaXiv:201709.00120v4

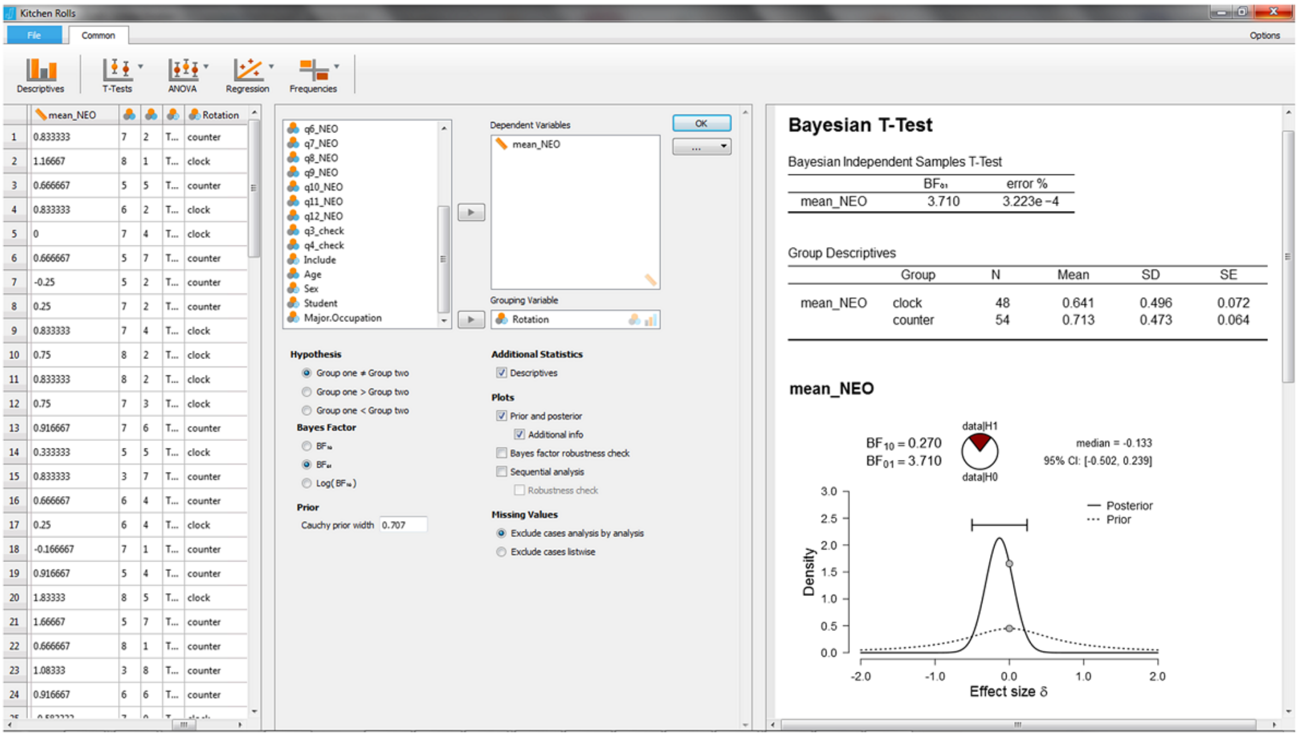

图 2 使用 JASP 进行贝叶斯独立样本 t 检验时的操作截屏。软件左侧是数据；中间为数据分析选项；右侧为结果输出。

通过这个初步的展示，我们可以了解到如何进行贝叶斯独立样本 t 检验的操作。接下来展示如何按照提前注册过的方法，对这批数据进行贝叶斯单侧独立样本 t 检验。由于描述性统计输出表明顺时针是组 1 而逆时针是组 2，我们将在”Hypothesis”的面板处勾选 “group 1 > group 2”，正如图 3 中间所示。

chinaXiv:201709.00120v4

chinaXiv:201709.00120v4

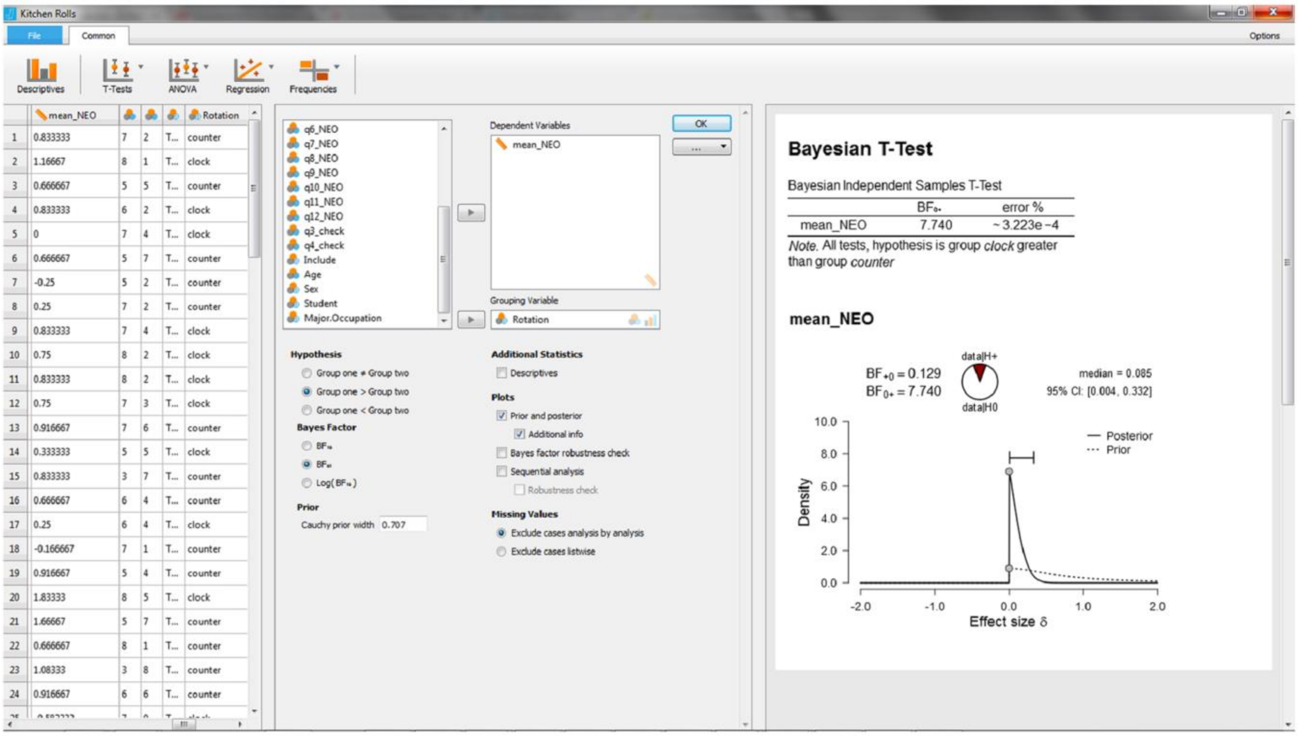

图 3 使用 JASP 对 Wagenmakers 等人(2015) (Wagenmakers, et al., 2015)数据进行贝叶斯单侧独立样本 t 检验的示意图。左侧是数据，中间为操作过程，右侧为结果输出。细节见文中的描述。

单侧检验的结果如图 3 右边部分所示。与预期的一致，如果观察到的效应是与假设相反，则这种使用单侧检验将先验知识整合到分析之中的做法，增加支持 H_0 的相对证据（也见 Matzke 等人(2015)），即贝叶斯因子 BF_{01} 从 3.71 增加到了 7.74，意味着观察到的数据在 H_0 下的可能是在 H_+ 可能性的 7.74 倍。

值得注意的是，在 H+下的后验分布是集中在 0 但不是没有负值（见图 3 右侧），与 H_+ 中的顺序限制是一致的。这一点与传统频率主义的单侧置信区间不同，传统方法的单尾置信区间为[-.23 +∞) [§]。虽然传统频率主义的区间在数学上是良好定义的（即，它包括了全部的不会被单尾的 $a = 0.05$ 显著性检验拒绝的值），但是大部分研究者会发现这个区间即不好理解也没有信息量(Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016)。

[§] 可以使用 R 语言中的 t.test 函数来得到 p 值的区间[-.23 +∞)。

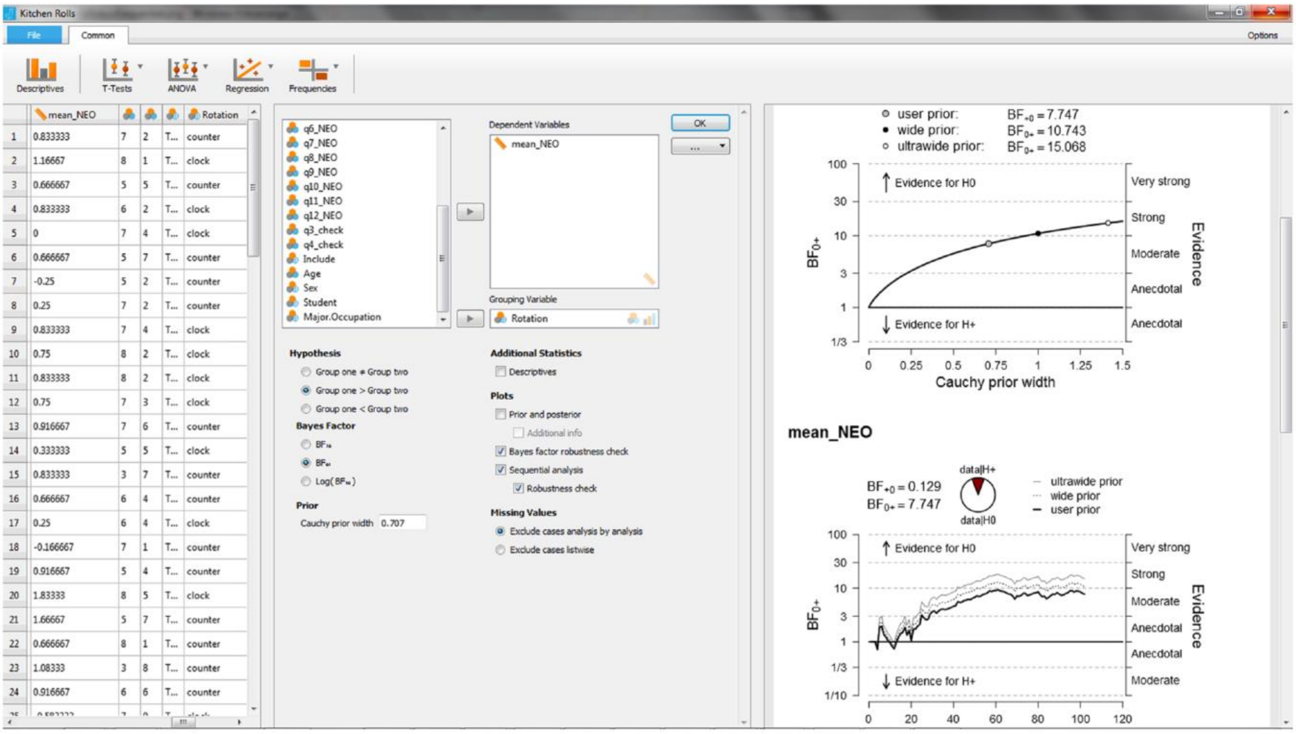

图 4 使用 JASP 进行贝叶斯因子的稳健性分析。

除了计算贝叶斯因子外，JASP 还可以进行稳健性分析（Bayesian robustness check），从而量化柯西先验分布参数 r 对贝叶斯因子的影响。如图 4 所示，选中“Bayes factor robustness check”的选项，这将得到图 4 右侧上面的图。从该图可以看到，当 Cauchy 先验的 r 为 0 时，H_0 与 H_+相同（$BF_{0+} = 1$），BF_{0+}随着 r 的增加而增加。在 JASP 的默认值 $r = 0.707$，贝叶斯因子 $BF_{0+} = 7.73$；而对于 Jeffrey 默认的 $r = 1$，贝叶斯因子 $BF_{0+} = 10.75$。因此，在一系列 r 的先验值中，当前数据显示了对 H_0 的中等到强的证据支持。

此外，还可以勾选图 4 中间的部分的“Sequential analysis”及其子选项“Robustness check”，进行序列分析。其结果见图 4 右侧下半部分的图。序列分析显示的是贝叶斯因子随着着抽样而变化的结果，也就是说，研究都可以在新数据收集到时对证据的积累进行监控和可视化。从图中可以看到，实际上 Wagenmakers 等人(2015)并未按预注册时的 $\gamma = 1$ 先验来计算 BF_{0+}并在 $BF_{0+} > 10$ 或者 $BF_{+0} > 10$ 时立刻停止收集数据: 在 55 个被试之后，虚线超过了 $BF_{0+} > 10$，但是数据仍然继续收集。在实践中，每隔几天检验一次贝叶斯因子，有助于了解贝叶斯因子

chinaXiv:201709.00120v4

chinaXiv:201709.00120v4

是否在某个时间点上超过预先决定的标准，并据此决定是否停止数据。

序列分析的一个优点是它可视化了贝叶斯因子在不同先验条件下的收敛过程，即贝叶斯因子在 log 尺度上差异开始稳定不变 (如, Bahadur & Bickel, 2009; Gronau & Wagenmakers, 2017)。在当前的例子中，当被试数量达到 35 时，不同先验下的贝叶斯因子开始出现收敛。要理解为什么在贝叶斯因子的 log 值的差异会在一些初步的观测数据之后不再变化，我们可以假定数据 y 包括两个部分 y1 和 y2，根据条件概率公式，$BF_{0+}(y) = BF_{0+}(y1) \times BF_{0+}(y2|y1)$。这个公式表明，贝叶斯因子并非是对不同数据进行盲目地相乘，实际上公式中的第二个因子——$BF_{0+}(y2|y1)$ —— 反映的是：当先验分布已经根据数据 y1 进行更新后，数据 y2 对贝叶斯因子再次更新 (Jeffreys, 1961, p. 333)。对这个公式进行 log 转换后，得到 $\log(BF_{0+}(y)) = \log(BF_{0+}(y1)) + \log(BF_{0+}(y2|y1))$。假定数据 y1 包括了足够的信息，不管 r 如何变化，通过 y1 更新后均得到大致相同的结果分布（在大部分情况下，这种情况很快就会出现）。而通过 y1 得到的这个后验分布，又变成了数据 y2 的先验分布，即成为 $\log(BF_{0+}(y2|y1))$的先验。在这种情况下，$\log(BF_{0+}(y2|y1))$的值大致相似（相似的先验分布，相同的数据）。因此，不一样的 r 值会让数据 y1 产生不同的后验分布，但当数据 y1 是足够的大后，使得 y1 的后验分布大致相似，此时 y2 再次对模型进行更新的大小也是相似，这就使得 $\log(BF_{0+}(y2|y1))$在不同的 r 下相似，产生收敛的现象。

3.3 如何报告贝叶斯因子结果

贝叶斯统计在目前的心理学研究中并不常见。虽然大部分杂志的编辑和审稿人会欣赏采用更加合理的统计手段，但是出于对贝叶斯方法的陌生，研究者使用贝叶斯因子时，需要提供相关的背景信息让编辑和审稿人了解这种背景。因此，除了报告贝叶斯因子的结果之外，还需要首先报告如下几点(Kruschke, 2014)。第一、选用贝叶斯因子的动机与原因，即为什么在

某个报告中使用贝叶斯因子而不是 NHST。如前所述，可以说明贝叶斯因子提供了更加丰富的信息，或者数据特点不满足 NHST 的前提假设（如在自然情境下收集的数据，无法判断数据收集的动机和实验假设）。第二、描述贝叶斯因子在模型比较中的基本逻辑。即，假定读者并不非常了解贝叶斯因子，简单地解释贝叶斯因子中模型比较的思想。第三、描述贝叶斯因子分析中的先验分布以及采用该先验的原因，先验分布应该或多或少对数据分析提供一些信息。第四、解释贝叶斯因子，将贝叶斯因子与研究中的理论或假设结合起来。

贝叶斯因子不使用统计显著，而是描述数据对假设的支持程度。例如，在 Wagenmakers 等 (2015)中，对 Jeffreys 默认先验下的贝叶斯因子结果进行如下描述：

“贝叶斯因子为 $BF_{01} = 10.76$，说明在（假定没有效应的）零假设下出现当前数据的可能性是在（假定存在效应的）备择假设下可能性的 10.76 倍。根据 Jeffreys (1961)提出的分类标准，这是较强的证据支持了零假设，即在顺时针和立逆时针转钟表指针的人在经验开放性（NEO）得分上没有差异。”

此外，使用贝叶斯因子进行分析时，还可以报告探索性的结果，如稳健性分布和序列分析的结果，这将进一步丰富结果，给其他研究者提供更加全面的信息。

4 总结与展望

近年来，科学研究的可重复问题备受关注(Baker, 2016; Begley & Ellis, 2012; Munafò et al., 2017)，在心理学(Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration, 2015)、神影像学(Poldrack et al., 2017; Zuo & Xing, 2014)尤其如此。而对 NHST 的过度依赖正是原因之一(Lindsay, 2015; 胡传鹏等, 2016)。因此，研究者们希望贝叶斯因子作为一种假设检验方法，能改变当前心理学研究中对 NHST 过度依赖的现状。当然，也有研究者提出了其他的方案，例如，将显著性的阈限下降到 0.005(Benjamin et al., 2017)或是采用模型比较的似然性比（likelihood ratio）(Etz, in press)的方法。但是值得注意的是，心理学研究重复失败的原因多种

chinaXiv:201709.00120v4

多样，仅改变统计方法不能让心理学的研究变得可重复。数据不开放以及研究过程不透明(Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Lindsay, 2015; Nosek et al., 2015)、对探索性分析与验证性分析不加区分(Kerr, 1998; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012)、以发表论文为核心的奖励体系(Nosek, Spies, & Motyl, 2012)等都可能是造成当前研究可重复率低的原因。因此从某种程度上来讲，对数据分析过程与数据结果保持开放与透明是关键的解决方案(如, (Poldrack & Gorgolewski, 2017; Zuo et al., 2014)。

即便如此，作为一种不同于传统 NHST 的方法，贝叶斯因子有助于研究者使用多种的方法对同一研究进行分析，从而得到准确的统计推断，得到更加接近真实的结论。需要指出的是，采用多种方法进行分析时，需要报告全部的分析过程和结果，而非选择最有利于自己结论的分析结果进行报告。

chinaXiv:201709.00120v4

4.1 贝叶斯因子的不足

贝叶斯因子是贝叶斯统计在假设检验方面的应用，而贝叶斯学派与频率学派统计的争议一直存在(Miller, 2011)。实际上，研究者指出贝叶斯因子也可能存在许多问题，充分了解这些反对的观点，将更加有利于我们在研究中合理地使用贝叶斯因子。

对贝叶斯因子最强烈的质疑来自于对其先验概率的设定，可能会认为先验概率过于主观、过于保守从而不容易出现较强的证据等(Wagenmakers, Marsman, et al., 2017)。也有研究者认为，默认的先验对小的效应不利。例如，Bem, Utts 和 Johnson (2011) 认为，Wagenmakers 等人(2011)对 Bem (2011)的数据进行重新分析时，采用了不合适的先验概率是他们未能得到与 Bem (2011)一致结论的原因。这种批评实质上是对贝叶斯因子的误用，即未将先验知识转化成为合适的先验概率(Hoijtink, van Kooten, & Hulsker, 2016)。有趣的是，只要研究者将自己使用的先验概率保持透明与公开，其他研究者可以采用交叉验证，从而起到充分探索的作用。

其次，也有研究者认为，贝叶斯因子没有考虑假阳性的问题。在 NHST 框架之下，研究者非常强调控制一类错误与二类错误的问题。例如心理学研究中一般将一类错误控制在 5%以内，因此显著性水平设置为 0.05。也正是需要控制一类错误，NHST 框架之下有许多方法用来调整阈值使其一类错误率不至于太高，例如多重比较校正的方法。而贝叶斯统计主要是为了不断地对证据的强度进行测量，其不考虑控制假阳性（即一类错误）的问题。因此，当研究者基于贝叶斯因子进行决策（效应是否存在）时，就可能犯下一类错误(Kruschke & Liddell, 2017a)。在实际的贝叶斯因子分析中，可以通过先验来解决多重比较的问题(Jeffreys, 1938; Scott & Berger, 2006, 2010)。例如，直接说明研究者预期假阳性率有多大(Stephens & Balding, 2009)。

还有研究者指出，基于估计的统计总是要比假设检验更优，因为估计本身将不确定性考虑进来。例如，Cumming (2014)建议使用效应量及其置信区间以替代 p 值。但是考虑到参数估计与假设检验在科研中均有其相应最适用的问题，因此贝叶斯因子无法直接与基于估计的频率主义学派统计进行比较。但是，贝叶斯统计中，也有基于估计的方法(Kruschke & Liddell, 2017b)。

最后，贝叶斯因子进行假设检验，本质上是证据的不断累积，而非得到二分的结论。因此，一次实验的结果可以被看作是试探性的，研究者可以继续收集数据或者进行重复实验(Ly, Etz, Marsman, & Wagenmakers, 2017)。

4.2 贝叶斯因子的应用前景

贝叶斯因子作为基于贝叶斯统计的假设检验方法，与 NHST 相比具有一些优势，其使得研究者可以直接检验数据是否支持零假设，不再受抽样意图和停止收集数据标准的影响，从而更加灵活地进行数据分析。这些优势可能帮助心理学家更好地在研究过程中进行决策，同

chinaXiv:201709.00120v4

时，贝叶斯因子的采用也可以促进研究者去更加深入地理解贝叶斯方向法的适用范围以及前提条件等(Depaoli & van de Schoot, 2017)。

JASP 的开发，使用贝叶斯因子的计算和解读变得更加简便，研究者即便没有很强的编程基础，也能够使用 JASP 地进行贝叶斯因子分析。这可能有助于推动研究者更加广泛地使用贝叶斯因子。此外，JASP 本身正在快速发展，其功能的深度和广度正在不断地扩大，新的方法和标准将不断地整合到软件之中，可能帮助研究者更科学地进行研究。

致谢：感谢清华大学社会科学学院心理学系张咪同学在本文撰写之初提供的帮助, 感谢两位匿名审稿人对本文提供的宝贵意见。

chinaXiv:201709.00120v4

参考文献：

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究的可重复性问题：从危机到契机. *心理科学进展, 24*(9), 1504–1518

骆大森. (2017). 心理学可重复性危机两种根源的评估. 心理与行为研究*, 15*(5), 577–586.

钟建军, Dienes, Z., 陈中永. (2017). 心理研究中引入贝叶斯统计推断的必要性、应用思路与领域. 心理科学 *, 40*(6), 1477–1482

Bahadur, R. R., & Bickel, P. J. (2009). An optimality property of Bayes' test statistics. *Lecture Notes-Monograph Series, 57*, 18–30.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 553*, 452–454.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407–425.

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology, 101*(4), 716–719.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., et al. (2017). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10.

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76*(2), 159–165.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2*(3), 317–335.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2 ed.). Hayward (CA): Institute of Mathematical Statistics.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan:

chinaXiv:201709.00120v4

A probabilistic programming language. *Journal of Statistical Software, 76*(1).

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. J. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at *AIMS Neuroscience* and beyond. *AIMS Neuroscience, 1*(2373-7972), 4–17.

Chen, X., Lu, B., & Yan, C.-G. (2018). Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Human Brain Mapping, 39*(1), 300–318.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods, 22*(2), 240–261.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. London, UK: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: which side are you on? *Perspectives on Psychological Science, 6*(3), 274–290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*(781).

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., et al. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin, 63*(6), 400–402.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193–242.

Etz, A. (in press). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57*(5), 153–169.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439–453.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., et al. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337–350.

Gronau, Q. F., & Wagenmakers, E.-J. (2017). Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics*, 1–10.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods, 12*(3), 179–185.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton, FL: Chapman & Hall/CRC.

Hoijtink, H., van Kooten, P., & Hulsker, K. (2016). Why Bayesian psychologists should change the way they use the Bayes factor. *Multivariate Behavioral Research, 51*(1), 2–10.

JASP Team. (2017). JASP (Version 0.8.2) [Computer software].

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society, 31*(2), 203–222.

Jeffreys, H. (1938). Significance tests when several degrees of freedom arise simultaneously.

chinaXiv:201709.00120v4

Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 165(921), 161–198.

Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences, 110*(48), 19313–19317.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: a "many labs" replication project. *Social Psychology, 45*(3), 142–152.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods, 10*(4), 477–493.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A Tutorial with R, JAGS, and Stan* (2 ed.). San Diego, CA: Academic Press/Elsevier.

Kruschke, J. K., & Liddell, T. M. (2017a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 1–23.

Kruschke, J. K., & Liddell, T. M. (2017b). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*.

Lakens, D. (2017). Equivalence tests: A practical primer for t-Tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics, 15*(1), 22–25.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*(12), 1827–1832.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28*(25), 3049–3067.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2017). Replication Bayes factors from evidence updating. *PsyArXiv*. Retrieved from https://osf.io/preprints/psyarxiv/u8m2s/

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2017). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology, 72*, 43–55.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016b). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19–32.

Marsman, M., & Wagenmakers, E.-J. (2016a). Bayesian benefits with JASP. *European Journal of Developmental Psychology, 14*(5), 545–555.

Marsman, M., & Wagenmakers, E.-J. (2016b). Three Insights from a Bayesian Interpretation of the One-Sided P Value. *Educational and Psychological Measurement, 77*(3), 529–539.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*(3), 679–690.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-

J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General, 144*(1), e1–e15.

Miller, G. (2011). ESP paper rekindles discussion about statistics. *Science, 331*(6015), 272–273.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*(1), 103–123.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406–419.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology, 53*(6), 530–546.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), 0021.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Paper presented at the Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003).

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafo, M. R., et al. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*(2), 115–126.

Poldrack, R. A., & Gorgolewski, K. J. (2017). OpenfMRI: Open sharing of task fMRI data. *NeuroImage, 144, Part B*, 259–261.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*(4), 682–689.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22*(2), 304–321.

Rouder, J. N., Speckman, P., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York, NY: W. H. Freeman and Company.

Schervish, M. J. (1996). P values: What they are and what they are not. *The American Statistician, 50*(3), 203–206.

Schlaifer, R., & Raiffa, H. (1961). Applied statistical decision theory

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2),

chinaXiv:201709.00120v4

322–339.

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference, 136*(7), 2144–2162.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statististics, 38*(5), 2587–2619.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ Values for testing precise null hypotheses. *The American Statistician, 55*(1), 62–71.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics, 10*(10), 681–690.

Stulp, G., Buunk, A. P., Verhulst, S., & Pollet, T. V. (2013). Tall claims? Sense and nonsense about the importance of height of US presidents. *The Leadership Quarterly, 24*(1), 159–171.

Topolinski, S., & Sparenberg, P. (2012). Turning the hands of time. *Social Psychological and Personality Science, 3*(3), 308–314.

van de Schoot, R., Winter, S., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods, 22*(2), 217–239.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54*(6), 491–498.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804.

Wagenmakers, E.-J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., et al. (2015). Turning the hands of time again: a purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology, 6*(494).

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology, 60*(3), 158–189.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2017). Bayesian Inference for Psychology. Part II: Example Applications with JASP. *Psychonomic Bulletin & Review*.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological Science Under Scrutiny* (pp. 123–138): John Wiley & Sons, Inc.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician, 70*(2), 129–133.

chinaXiv:201709.00120v4

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison ssing 855 t tests. *Perspectives on Psychological Science, 6*(3), 291–298.

Zhu, J., Chen, J., Hu, W., & Zhang, B. (2017). Big Learning with Bayesian methods. *National Science Review, 4*(4), 627–651.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance*. Ann Arbor: University of Michigan Press.

Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., et al. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Nature Scientific Data, 1*, 140049.

Zuo, X.-N., & Xing, X.-X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: A systems neuroscience perspective. *Neuroscience & Biobehavioral Reviews, 45*, 100–118.

chinaXiv:201709.00120v4

The Bayes Factor and Its Implementation in JASP: A Practical Primer

Chuan-Peng HU [1, 2]; Xiang-Zhen KONG [3]; Eric-Jan WAGENMAKERS [4], Alexander LY [4], Kaiping PENG [1]

(1 Department of Psychology, School of Social Science, Tsinghua University, Beijing, China. 10004;

2 Neuroimaging Center, Johannes Gutenberg University Medical Center, 55131 Mainz, Germany

3 Language and Genetics Department, Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands;

4 Department of Psychological Methods, University of Amsterdam, 1018 VZ Amsterdam, The Netherlands)

chinaXiv:201709.00120v4

Abstract: Statistical inference plays a critical role in modern scientific research, however, the dominant method for statistical inference in science, null hypothesis significance testing (NHST), is often misunderstood and misused, which leads to unreproducible findings. To address this issue, researchers propose to adopt the Bayes factor as an alternative to NHST. The Bayes factor is a principled Bayesian tool for model selection and hypothesis testing, and can be interpreted as the strength for both the null hypothesis H_0 and the alternative hypothesis H_1 based on the current data. Compared to NHST, the Bayes factor has the following advantages: it quantifies the evidence that the data provide for both the H_0 and the H_1, it is not "violently biased" against H_0, it allows one to monitor the evidence as the data accumulate, and it does not depend on sampling plans. Importantly, the recently developed open software JASP makes the calculation of Bayes factor accessible for most researchers in psychology, as we demonstrated for the *t*-test. Given these advantages, adopting the Bayes factor will improve psychological researchers' statistical inferences. Nevertheless, to make the

analysis more reproducible, researchers should keep their data analysis transparent and open.

Keywords: Bayes factor, Bayesian statistics, Frequentist, NHST, JASP

chinaXiv:201709.00120v4